LAMP-TR-125 August 2005
CS-TR-4754
UMIACS-TR-2005-57

A Paraphrase-Based Approach to Machine Translation Evaluation

Grazia Russo-Lassner[1,3], Jimmy Lin[2,3] and Philip Resnik[1,3]
[1]Department of Linguistics
[2]College of Information Studies
[3]Institute for Advanced Computer Studies
University of Maryland
College Park, MD 20742
{glassner, jimmylin, resnik}@umiacs.umd.edu

Abstract

We propose a novel approach to automatic machine translation evaluation based on paraphrase identification. The quality of machine-generated output can be viewed as the extent to which the conveyed meaning matches the semantics of reference translations, independent of lexical and syntactic divergences. This idea is implemented in linear regression models that attempt to capture human judgments of adequacy and fluency, based on features that have previously been shown to be effective for paraphrase identification. We evaluated our model using the output of three different MT systems from the 2004 NIST Arabic-to-English MT evaluation. Results show that models employing paraphrase-based features correlate better with human judgments than models based purely on existing automatic MT metrics.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
AUG 2005		**00-00-2005 to 00-00-2005**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
A Paraphrase-Based Approach to Machine Translation Evaluation	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **University of Maryland,Institute for Advanced Computer Studies,College Park,MD,20742**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
The original document contains color images.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**		**11**	

1 Introduction

While human evaluation of machine translation (MT) output remains the most reliable method to assess translation quality, it is a costly and time consuming process. The development of automatic MT evaluation metrics enables the rapid assessment of system output. By providing immediate feedback on the effectiveness of various techniques, these metrics have guided machine translation research and have facilitated rapid advances in the state of the art. In addition, automatic evaluation metrics are useful in comparing the performance of multiple MT systems on a given translation task, as demonstrated by the DARPA TIDES research program. Since automatic evaluation metrics are meant to serve as a surrogate for human judgments, their quality is determined by how well they correlate with assessors' preferences and how accurately they predicts human judgments.

Although current methods for automatically evaluating MT output do not require humans to assess individual system output, humans are nevertheless needed to generate a number of "reference translations". The quality of machine-generated translations is determined by automatically comparing system output against these references. Despite numerous variations, which we will discuss in the next section, all current automatic evaluation metrics are based on the simple idea of matching substrings (i.e., *n*-grams) from machine output with substrings from the reference translations. This substring matching approach has obvious drawbacks: it does not account for combinations of lexical and syntactic differences that might occur between a perfectly fluent and accurately-translated machine output and a human reference translation (beyond variations already captured by the different reference translations themselves). Moreover, the set of human reference translations is unlikely to be an exhaustive inventory of "good translations" for any given foreign language sentence. Therefore, it would be highly desirable to have an MT evaluation metric capable of automatically determining equivalences in meaning without relying on exact substring matches.

We propose a novel approach to automatic machine translation evaluation based on paraphrase identification. The quality of machine-generated output can be viewed as the extent to which the conveyed meaning matches the semantics of the reference translations, independent of substrings they may share. In short, all paraphrases of human-generated references should be considered "good" translations. We have implemented this idea in a statistical model of human preferences that combines features from existing automatic evaluation metrics with features that have proven to be useful in the paraphrase identification task. Results show that exploiting paraphrase identification techniques results in a statistically significant improvement in correlation with human judgments at the sentence level, measured against baselines that use only existing automatic metrics.

Some might argue that, at current levels of MT system performance, greater fidelity to human judgments is beside the point — perhaps systems still have ample room for improvement before there is added value in features not based on substring matches or in models that go significantly beyond *n*-gram overlap with reference translations. However, recent developments in MT evaluation point to a greater emphasis on human preferences. In the new DARPA GALE program (DARPA, 2005), MT evaluation will be performed by measuring the translation error rate (TER, Snover et al., 2005) between a

system's output hypothesis H and the string H' created by a human who corrects the machine output to turn it into an acceptable translation. TER is a variant of string edit distance inspired by the idea of modeling how a human would turn system output into the "closest" acceptable translation. While this human-in-the-loop use of TER is a positive development, allowing for variations not adequately captured by substring matches against reference translations, it introduces a dilemma: systems can no longer be tested frequently in development or automatically optimized using the same objective function that will be used for their evaluation. As a remedy, we believe that it is important to develop models of translation quality that can be computed automatically, like current *n*-gram based metrics, but which better approximate what will be taking place during the human-in-the-loop evaluation task. This task can be conceptualized, abstractly, as an iterative process that includes formulating a representation of the correct translation, assessing the extent to which the system hypothesis conveys the same meaning, and modifying the hypothesis so that it better conveys the intended meaning. Since a central part of this task closely resembles judging the "goodness" of paraphrase, we hypothesize that our approach, which incorporates features known to be useful for identifying paraphrases, will lead to better automatic metrics.

2 Related work

MT evaluation is a complex task; no single metric, manual or automatic, can adequately capture all factors that contribute to a good translation. Evaluation criteria might depend on translation domain, task, and users involved. FEMTI (Framework for the Evaluation of Machine Translation), a recent effort to group various existing metrics into a unifying framework, proposes two taxonomies, one relating an MT system's use context to a quality model, and the other relating the quality model to appropriate metrics (Hovy et al., 2002).

Despite the challenges in quantifying a "good" translation, researchers have made substantial progress in automatic MT evaluation. The most successful metrics currently in use are based on substring (i.e., *n*-gram) matches between machine output and one or more human-generated reference translations at the sentence level. This simple idea is implemented in the BLEU metric (Papineni et al., 2002) and the closely-related NIST metric (Doddington, 2002), both of which have been shown to correlate with human judgment when applied on multiple-sentence testsets (entire documents or sets of documents). Despite widespread adoption, these metrics present several drawbacks, which MT researchers have tried to address. For instance, the team at the 2003 Johns Hopkins Summer Workshop on Syntax for Statistical MT noticed that BLEU is insensitive to syntactic changes (Och et al., 2003); the METEOR metric (Banerjee et al., 2005), discussed further below, was developed to specifically address BLEU's emphasis on *n*-gram precision, which does not appropriately measure the degree to which a machine-generated translation captures the entire content of the source sentence. Another weakness of the abovementioned MT evaluation metrics is that they do not correlate well with human judgment at the sentence level, despite correlations over large test sets (Blatz et al., 2003; Och et al., 2003; Kuleska et al., 2004). Clearly, automatic metrics with high sentence-level correlation are desirable because they provide a finer-grained assessment of translation quality. In particular, they can guide MT system development by offering feedback on sentences that are particularly challenging.

Evaluation of machine-translated output has been viewed as a sentence-level classification problem. Blatz et al. (2004) employ a feature-based machine learning approach to assess the confidence estimate of MT output both at the sentence level and at the word level. Their data consists of (**x**, *c*) pairs, where **x** is a feature vector representing the source/target-candidate translation and *c* is a correctness score. The correctness score is set to different thresholds of word error rate or the NIST score, depending on the experiment; the features are derived either from the base MT system (such as translation model probability, number of jumps in word-to-word alignment, language model, IBM Model 1 translation probabilities) or are derived from external sources (such as a semantic similarity metric based on WordNet). While sharing a supervised feature-based machine learning approach, this work differs from that of Blatz et al. in the choice of features; we utilize linguistically-motivated features useful for paraphrase identification, and model human translation quality judgments directly.

Similarly, Kuceska et al. (2004) view MT output evaluation as a classification problem. In an attempt to eliminate the need for human judgments, they train a support vector machine (SVM) to distinguish machine-generated translations from human translations. The continuous "confidence" probability generated by the SVM is used as a measure of translation quality. Kuceska et al. use features based on existing evaluation metrics: *n*-gram precision, ratio of hypothesis length to reference length, word error rate, and position-independent word error rate.

An attempt at going beyond *n*-gram matching is made with the METEOR metric (Banerjee et al., 2005), which is based on a word-to-word alignment between the machine-generated translation and the reference translation. This metric assigns a score equal to the harmonic mean of unigram precision (that is, the proportion of matched *n*-grams out of the total number of *n*-grams in *the evaluated translation*) and unigram recall (that is, the proportion of matched *n*-grams out of the total number of *n*-grams in *the reference translation*). METEOR also includes a fragmentation penalty that accounts for how well-ordered the matched unigrams of the machine translation are with respect to the reference. The alignment between machine translation and reference translation is obtained through mapping modules that apply sequentially, linking unigrams that have not been mapped by the preceding modules: the 'exact' module maps words that are exactly the same; the 'porter-stem' module links words that share the same stem; the 'WordNet synonymy' module maps unigrams that are synonyms of each other.

The mapping modules in METEOR are similar to some of the features used in our study. However, we go beyond word-to-word mappings to features that match multiple content words in the system hypothesis and the reference translation (approximating dependency relations). Moreover, we do not impose any brevity penalty, because multiple strings can share the same meaning, but be of different lengths.

A syntax-based approach to MT evaluation is explored in Liu et al. (2005); they propose two metrics, one based on the number of subtrees common to a hypothesis and a reference translation, and the second one computing the fraction of head-word chains occurring in both machine output and reference translation. Their technique is shown to improve correlations with fluency scores. Our work also captures dependency relations through the "composite" features, as shown in Section 3.2, but does not require the existence of a parser.

Recently, Tate et al. (2005) have developed a regression model for task-based performance in an information extraction task, using an automatic intrinsic metric derived from BLEU; the approach has the potential to replace costly task-based evaluations by taking advantage of intrinsic measures that can be computed automatically. Their work represents a potential consumer for improved intrinsic metrics of the kind we propose here.

3 Methods

Our basic approach for modeling human translation quality employs a linear regression model trained on (**x**, *h*) pairs, where **x** is a feature vector representing corresponding sentence pairs (between machine-translated output and human references), and *h* is a human evaluation score (see below). A linear regression model was appropriate for this task because the dependent variable has a normal distribution and can be treated as continuous. In addition, each feature value in our data set has a roughly linear relation with *h*. To evaluate each model, we compute its Pearson's *r* correlation with true human judgments.

3.1 Data

Data for our experiments consist of 347 sentences from the DARPA/TIDES 2004 Arabic-to-English MT evaluation testsets, translated by three different MT systems (identified as arf, ari, arp). These three systems were chosen because they represent high, medium, and low performing translation systems as measured by BLEU in the NIST 2004 evaluation. For each of the system's output, two human annotators have manually assigned a score from 1 to 5 for fluency and adequacy; we used the average of the two annotators' scores. Thus, we have a total of 1041 training samples. There are four reference translations available for this data set. For this study, we created three separate models that attempt to capture fluency, adequacy, and the average of the two values.

3.2 Features

Because our approach views automatic machine translation evaluation as paraphrase identification, we employed features previously shown to be useful for that task. These features attempt to go beyond simple *n*-gram matching to account for lexical and syntactical variations between machine and reference output. In particular, we have implemented the features proposed in (Hatzivassiloglou et al., 1999; Hatzivassiloglou et al., 2001), which were used to detect semantic similarity between sentence-sized text segments. In their work, a distinction is made between features that compare single terms from two sentences, called 'primitive' features, and those that match word pairs to word pairs, called 'composite' features. The following 'primitive' features are based on word-to-word correspondences between a machine-translated output and a reference translation:

- **Stemmed words co-occurrence (Wrd):** matching of tokens generated by the Porter stemmer.
- **Noun phrases (NP)**: matching of unstemmed noun phrase heads. Noun phrase bracketing is accomplished by matching regular expression patterns of part-of-speech tags over output generated by the Brill tagger.

- **WordNet synsets (WnSyn)**: matching of words that appear in the same WordNet 2.1 synset (Miller et al., 1990; Fellbaum, 1999). No word sense disambiguation is performed.
- **Verbs semantic classes (Verb)**: matching of verbs that share the same semantic class, as defined by Levin (1993).
- **Proper names of person, place, organization (PNP)**: matching of proper names based on part-of-speech tags.

The 'composite' features establish correspondences between word pairs from a machine translation and a reference translation, where the word pairs are identified by the 'primitive' features. These 'composite' features roughly capture dependency relations:

- **Order (Order):** matches pairs of 'primitive' features if they occur in the same relative order in both sentences.
- **Distance (D2, D3, D4, D5)**: matches pairs of 'primitive' features if they occur within a window of 2-5 words.

The actual values assigned to the 'primitive' and 'composite' features correspond to the number of matches between the machine output and the reference translation; not all reference translations were taken into account in our current experiments, but only the one which has the least number of translation error according to TER. To avoid bias in favor of longer segments, each feature value is normalized by the length of the sentences. Following (Hatzivassiloglou et al., 1999), the feature values of two sentences A and B are divided by

$$\sqrt{length(A) \times length(B)}\,.$$

In our experiments, we also included features based on existing MT evaluation metrics:

- **BLEU score**: the most commonly-used automatic MT evaluation metric today; see Section 2 for more details.
- **TER (translation error rate)**: this recently-introduced metric (Snover et al., 2005) is based on the number of edits (insertions, deletions, substitutions, shifts) it takes a human to convert the system output into a correct translation. Unlike Word Error Rate (WER), another automatic MT evaluation metric (Vidal, 1997; Tillmann et al., 1997), TER allows word shifts: it treats shifts of contiguous multi-word sequences as a single operation. Similarly to NIST and BLEU, TER is defined for multiple references.
- **METEOR score**: an automatic MT evaluation metric based on a combination of unigram-precision and unigram-recall with the reference translations (Banerjee et al., 2005), as discussed in Section 2.

3.3 Models

We constructed three separate models, one for human-rated adequacy, one for fluency, and one for the average of both. For each model, we experimented with the following feature combinations: BLEU ± paraphrase, TER ± paraphrase, METEOR ± paraphrase, BLEU + TER ± paraphrase, BLEU + TER + METEOR ± paraphrase, where "paraphrase" indicates that the paraphrase features were included in the regression model.

Our models were evaluated by measuring the sentence-level Pearson's *r* correlation between model output and true human judgments over the entire data set. In computing correlations, it is acceptable not to have a training/test division because we are currently more concerned about modeling human preferences instead of developing a predictive MT evaluation metric.

In addition to the more widely known Fisher's *z* test for assessing the statistical significance of correlation differences, we employ Meng, Rosenthal and Rubin's *z* transformation (MRR) (Meng et al., 1992). Fisher's *z* test is only appropriate when comparing two independent samples, a condition not met here: we are comparing the correlation between one pair of variables and a second, overlapping pair of variables.[1] Moreover, our data come from the same set of translations for all variables. Fisher's *z* results are reported only to maintain consistency with related work in the literature.

4 Results

Table 1 shows the Pearson's *r* correlation coefficients of each model with human judgments of adequacy, fluency, and an average of both values. Relative improvements in correlation over a BLEU baseline are shown, as well as the statistical significance of the gains.

Model	Adequacy	Fluency	Average
B	.495	.392	.477
T	.542 (+9.5%) ‡	.440 (+12.2%) ‡	.529 (+10.9%) ‡
M	.627 (+26.7%) †‡	.491 (+25.3%) †‡	.602 (+26.2%) †‡
BP	.609 (+23.0%) †‡	.488 (+24.5%) †‡	.590 (+23.7%) †‡
TP	.617 (+24.6%) †‡	.490 (+25.0%) †‡	.600 (+25.8%) †‡
MP	.637 (+28.7%) †‡	.506 (+29.1%) †‡	.616 (+29.1%) †‡
BT	.554 (+11.9%) ‡	.446 (+13.8%) ‡	.538 (+12.8%) ‡
BTP	.617 (+24.6%) †‡	.499 (+27.3%) †‡	.600 (+25.8%) †‡
BTM	.640 (+29.3%) †‡	.508 (+29.6%) †‡	.618 (+29.6%) †‡
BTMP	.647 (+30.7%) †‡	.513 (+30.9%) †‡	.624 (+30.8%) †‡

Table 1. Pearsons' *r* correlation and relative improvement over BLEU baseline for each model of human judgments (adequacy, fluency, and average of both). † indicates significance at the 99% level by the Fisher's *z* test, ‡ indicates significance by the MRR *z* test. (B = BLEU, T = TER, M = METEOR, P = paraphrase)

[1] An example of overlapping pairs of variables is the correlation between BLEU and human judgment (HJ) and the correlation between TER and HJ, where HJ is common to the two variable pairs.

A comparison of models with and without paraphrase-based features is found in Table 2. The introduction of paraphrase-based features results in a statistically-significant improvements in correlation for BLEU and TER baseline models. Based on this current data set, we have insufficient evidence to conclude that the paraphrase-based features have any effect on the METEOR models or models that incorporate BLEU, TER, and METEOR. A more thorough evaluation with a larger data sample and cross-validation is currently being conducted, together with an error analysis to determine how the paraphrase-based features help and where they fail to enhance the model with respect to the baseline metrics.

	Adequacy	**Fluency**	**Average**
B vs. BP	+23.0% †‡	+24.5% †‡	+23.7% †‡
T vs. TP	+13.8% †‡	+11.4% ‡	+13.4% †‡
M vs. MP	+1.6%	+3.1% *	+2.3%
BTM vs. BTMP	+1.1%	+1.0%	+1.0%

Table 2. Statistical significance of correlation improvements for all paraphrase-based models with respect to their respective baselines. † indicates significance at the 99% level by the Fisher's *z* test, * indicates significance at 95% level by the MRR *z* test, ‡ indicates significance at the 99% level by the MRR *z* test.

To assess the contribution of each feature, we introduced each independent variable into the model stepwise, for all experiments reported. This means that at each step, of the independent variables that have not been added to the model, the one with the smallest F statistics is added; and at each step, of the variables already in the model, those with sufficiently large F statistics are removed. This method terminates when no more variables are eligible for inclusion or removal.

Table 3 presents a list of the paraphrase-based features that have contributed in a statistically significant way to their respective models. To increase legibility of the table, we have replaced the names of the contributing features with digits as follows: BLEU = 1; TER = 2; METEOR = 3, Wrd = 4, WnSyn = 5, O = 6; D2 = 7.

Model	**Adequacy**	**Fluency**	**Average**
BP	1 4 5	1 7 6 5	1 4 5 6 7
TP	2 4 5	2 7 6 5	2 5 7 6
MP	3 4 5	3 7 6	3 7 5 6
BTP	2 4 5	2 5 7 6	2 5 7 6
BTMP	2 3 5	2 3 5	2 3 5

Table 3. The features that have been statistically significant in building their respective models. (BLEU = 1; TER = 2; METEOR = 3, Wrd = 4, WnSyn = 5, O = 6; D2 = 7)

The feature contributions make intuitive sense. Generally speaking, correlation with translation adequacy is best when the model provides for more flexible word matching (found in METEOR and in paraphrase features involving stemming and synonymy) and

matches of higher order *n*-grams (BLEU, TER). Features reflecting locality and the preservation of relative order (TER, D2, Order) improve predictions of fluency.

5 Ongoing Work

The results obtained thus far suggest that there is promise in the paraphrase-based approach to machine translation evaluation. Further experimentation is needed to establish that improved correlations with human judgments are found when the model is used predictively, i.e., in scenarios where regression coefficients are fixed and the model is then applied to previously unseen data. Our immediate research goals include the following:

- **Error analysis and collection of larger data sample.** We plan to conduct an error analysis in order to identify where our models improve upon the baselines and where they fail to bring about improvements, and to analyze in more detail the effects of each feature on the correlation with human judgments. Larger and more varied datasets are also needed, since experiments in this study were limited to the output of three MT systems, all translation from the same foreign language, and to a rather small set of translations (347 per MT system).
- **Less direct modeling of human judgments.** An approach based on directly modeling human judgments has some limitations. First, it requires a large training set of human-evaluated hypothesis and reference translations; second, such a set of training data would have to be updated often to reflect the changing population of MT outputs. Therefore, we plan to explore (a) embedding paraphrase-detection techniques into a framework similar to that of Kulesza et al. (2004), in which a learning model is trained to distinguish between human-produced and machine-produced translations, and requires a one time start-up cost to assemble a large set of manually evaluated MT outputs; and (b) implementing paraphrase-based modules in a fashion similar to the approach of (Banerjee et al., 2005).
- **An empirical study of paraphrase phenomena in MT system output.** It would be extremely informative to perform a quantitative analysis of paraphrase-related phenomena (informed by characterizations of paraphrase types, e.g. Dorr et al., 2004) in the output of MT systems — for instance, to investigate which phenomena tend to reflect mistranslations versus valid paraphrases when found in MT output.

6 Conclusions

This work proposes a novel paraphrase-based approach to automatic machine translation evaluation. The idea is relatively simple: all machine-generated translations that are paraphrases of human references should be considered "good". The task of MT evaluation, therefore, becomes one of paraphrase identification. Our experiments show that this is a promising approach to more realistic MT evaluation than existing techniques based on substring matches. We demonstrate that a regression model incorporating paraphrase features can improve on standard MT evaluation baselines such as BLEU, METEOR, and TER in correlations with human judgments of adequacy and fluency.

Acknowledgements

This study was supported in part by Department of Defense contract RD-02-5700 and ONR MURI Contract FCPO.810548265. We are grateful to Matt Snover and Bonnie Dorr for making the TER implementation available and for valuable assistance with data preparation, and to Chip Denman for valuable help with the statistical side of this work.

References

Banerjee, S., and A. Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. *Proceedings of the ACL 2005 Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*.

DARPA Information Processing Technology Office. 2005. Global Autonomous Language Exploitation. http://www.darpa.mil/ipto/programs/gale/

Dorr, B. J., R. Green, L. Levin, O. Rambow, D. Farwell, N. Habash, S. Helmreich, E. Hovy, K. J. Miller, T. Mitamura, F. Reeder, and A. Siddharthan. 2004. Semantic Annotation and Lexico-Syntactic Paraphrase. *Proceedings of the Workshop on Building Lexical Resources from Semantically Annotated Corpora*, LREC, Portugal.

Fellbaum, C. ed. 1999. *WordNet: an electronic lexical database*. Cambridge, MA: The MIT Press.

Hatzivassiloglou, V., J. Klavans, and E. Eskin. 1999. Detecting Text Similarity Over Short Passages: Exploring Linguistic Feature Combinations via Machine Learning. *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-1999)*.

Hatzivassiloglou, V., J. Klavans, M. Holcombe, R. Barzilay, and M. Kan. 2001. SimFinder: A Flexible Clustering Tool for Summarization. *Proceedings of the NAACL 2001 Automatic Summarization Workshop*.

Hovy, E., M. King, and A. Popescu-Belis. 2002. Principles of Context-Based Machine Translation Evaluation. *Machine Translation*, 16:1-33.

Gamon, M., A. Aue, and M. Smets. 2005. Sentence-level MT Evaluation Without Reference Translations: Beyond Language Modeling. *Proceedings of the 10th Annual EAMT Conference*.

Kuleska, A., and S. M. Shieber. 2004. A Learning Approach to Improving Sentence-Level MT Evaluation. *Proceedings of the 10th International Conference on Theoretical and Methodological Issues in Machine Translation*.

Levin, B. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. Chicago, Illinois: University of Chicago Press.

Liu, D., and D. Gildea. 2005. Syntactic Features for Evaluation of Machine Translation. *Proceedings of the ACL 2005 Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization.*

Meng, X., R. Rosenthal, and D. Rubin. 1992. Comparing Correlated Correlation Coefficients. *Psychological Bulletin*, 111:172-175.

Miller, G. A., R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller. 1990. Introduction to WordNet: An On-Line Lexical Database. *International Journal of Lexicography*, 3(4):235-312.

Snover, M., B. J. Dorr, R. Schwartz, J. Makhoul, L. Micciulla, and R. Weischedel. 2005. A Study of Translation Error Rate with Targeted Human Annotation. Technical Report LAMP-TR-126, CS-TR-4755, UMIACS-TR-2005-58, University of Maryland, College Park.

Tate, C., C. Voss, B. J. Dorr, and E. Slud. 2005. Toward a Predictive Statistical Model of Task-based Performance Using Automatic MT Evaluation Metrics. *Proceedings of the Association for Computational Linguistics Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization,* Ann Arbor, MI.

Vidal, E. 1997. Finite-State Speech-to-Speech Translation. *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*.

Tillmann, C., S. Vogel, H. Ney, H. Sawaf, and A. Zubiaga. 1997. Accelerated DP Based Search for Statistical Translation. *Proceedings of the 5th European Conference on Speech Communication and Technology*.